Am. J. Hum. Genet. 66:1689–1692, 2000

Report

DNA Pooling in Mutation Detection with Reference to Sequence Analysis

Christopher I. Amos, Marsha L. Frazier, and Wenfu Wang

Departments of Epidemiology and Biomathematics, The University of Texas M. D. Anderson Cancer Center, Houston

We discuss pooling methods of mutation detection for identifying rare mutations. We provide mathematical formulae for obtaining the optimal pool size as a function of the mutation frequency in the study population and the specificity of the test. The optimal pool size depends strongly on the specificity of the test. With a test that has 99% specificity, pooling can reduce the number of tests that need to be performed by 80%, whereas, with a test with 95% specificity, pooling reduces the number of samples that must be tested by only 50%. We used the software PHRED to call mutations after sequencing of pooled samples with known STK11 mutations. We found that, when the area under the curve for the less prominent peak was used to call mutations, we were able to pool pairs of samples and correctly identify mutations. Pooling of three samples did not lead to an adequately specific test for the basic automated allele-calling procedures that we used. We discuss methods by which the specificity may be improved to permit pooling of three or more samples when testing for mutations by sequencing.

Pooling strategies have been advocated for genetic-linkage identification (Churchill et al. 1993; Sheffield et al. 1995), detection of clones for physical-mapping studies (Barillot et al. 1991; Bruno et al. 1995), and association studies (Daniels et al. 1998; Shaw et al. 1998) but have not been widely employed for mutation detection in individual patients. Nevertheless, the cost for mutation detection for genes such as BRCA1 and BRCA2 and the DNA mismatch–repair genes, hMLH1 and hMSH2, can be prohibitive. A typical mutation-detection protocol requires that for each individual to be tested, each exon—or, possibly, a few closely located exons—is PCR amplified and then assayed. BRCA1 mutations are among the more common major genes causing familial illness. Nevertheless, population estimates for the prevalence of BRCA1 or BRCA2 mutations range from <0.3% among non-Jewish whites (Claus et al. 1991), to ~2% for Ashkenazim (Hartge et al. 1999). Furthermore, mutations in BRCA1 or BRCA2 (and in most other cancer-predisposing loci) are scattered throughout the coding region for most populations, so that the probability that any particular amplified segment contains a mutation is much lower than the probability for the entire gene. Except in some special populations, common mutations of cancer-predisposing genes do not exist. The rarity of mutations within exons of these commonly studied genes further reinforces the need to develop DNA-pooling strategies to detect mutations more efficiently.

A major issue in single-nucleotide-polymorphism studies is identification of polymorphisms through resequencing of already cloned genes (Mohrenweiser and Jones 1998). For these studies, the targeted gene frequency is generally on the order of ⩾10% per exon, and pooling is not likely to be effective during the current period in which common alleles are sought. However, if future studies seek to identify unusual polymorphisms (Taillon-Miller et al. 1999), then resequencing efforts including larger numbers of subjects may benefit from some of the design issues we describe here.

A limiting factor in the use of pooling strategies is the sensitivity of the assay. By sensitivity, we mean the probability to detect a mutation given that the mutation is present in some member of the DNA pool. Data concerning the sensitivity of mutation detection methods in pooled samples is not available for the frequently used methods such as direct sequencing or single stranded conformational polymorphism analysis. However, for detection of mutations using multiplex single nucleotide primer extension, pooling of 10 or 20 samples led to an

Received April 20, 1999; accepted for publication February 9, 2000; electronically published March 24, 1999.

Address for correspondence and reprints: Dr. Christopher Amos, Departments of Epidemiology and Biomathematics, 1515 Holcombe Boulevard, Box 189, Houston, TX 77030. E-mail: camos@request.mdacc.tmc.edu

0002-9297/2000/6605-0024$02.00

Table 1

Mutation Frequency versus the Optimal Pooling Strategy and Average Sample Size Required with Pooling, as a Percentage of the Size Needed without Pooling

Mutation Frequency Per Exon (π)	Optimal Pool Size (p)	Average Optimal Sample Size with Pooling (y/n)	Optimal Pool Size with 5% False-Positive Results	Average Sample Size with Pooling with 5% False-Positive Results
.4	1	No improvement	1	No improvement
.2	1	No improvement	1	No improvement
.1	2	88.0%	2	96.1%
.05	2	69.5%	2	78.5%
.01	4	40.8%	3	56.8%
.005	5	32.4%	4	52.5%
.001	7	18.9%	4	46.5%
.0005	10	15.0%	4	45.8%
.0001	17	8.8%	4	45.2%

estimated 100% sensitivity, whereas, for pools of 30 samples, the sensitivity dropped to 80% (Krook et al. 1992). Coolbaugh-Murphy et al. (1999) found that for detecting microsatellite genotypes that pools of $\leqslant 5$ genome equivalents provided a test with adequate sensitivity. In this study, samples were pooled prior to the PCR amplification. For direct sequencing, data are not available concerning the sensitivity of pooled samples for mutation detection, and we provide results from initial studies on this issue. Finally, the mismatch amplification mutation assay is sensitive for detecting specific mutations in large pools (on the order of 1 mutation in 10^5 samples) so that multistage pooling may be feasible with this assay, but the sensitivity and specificity were not clearly presented (Chen and Zarbl 1997). In this letter, we are assuming that the maximal pool size to ensure 100% sensitivity for the assay is small (<10) so that multistage pooling is not appropriate. The statistical approach is applicable for any type of assay with relatively small pools but we have specifically studied issues related to sequence analysis.

Suppose that samples are obtained from n subjects and that one wants to identify the optimal number of samples to pool, r. Assume that the probability of detecting a mutation in a pool of size r is given by γ. Let the probability of a mutation in a single exon (or other unit being studied) be π. The number of mutation-detection runs that must be completed without pooling is just n. With single-stage pooling, the expected number of runs to complete the mutation detection, y, is given by

$$y = \frac{n}{r} + n\gamma r \ .$$

The sample is first organized into n/r pools. We expect that $n\gamma$ runs will show mutations and that, for each of these pools, r samples will have to be resequenced to identify the sample(s) that contain the mutation(s). Any given pool that is found to have a mutation could contain one or more mutations, so every sample in the pool needs to be sequenced. Now, $\gamma = 1 - (1 - \pi)^r$, because the probability of detecting at least one mutation in a pool is 1 minus the probability of detecting no mutations in the r independent pools of the sample. Thus, we need only minimize the equation

$$y = n\left\{\frac{1}{r} + r[1 - (1 - \pi)^r]\right\}$$

over r to obtain an optimal pool size. Differentiating y with respect to r, the resulting equation to solve is:

$$0 = 1 - \frac{1}{r^2} - (1 - \pi)^r[1 + r\log_e(1 - \pi)]$$

This formula does not have a simple solution. However, some values minimizing this expression have been provided in table 1, along with the required expected number of samples that will have to be performed. Evaluation of the second derivative confirms that, for small values of π, the function is minimized (results not shown). When studying sex-linked traits, π can be taken to be the gene frequency of the disease allele, while for autosomal traits, if p is the allele frequency, then $\pi = 2p(1 - p) + 2p^2$.

False-positive findings increase the number of samples that will have to be assayed. If we let β be the probability of a false-positive result (and $1 - \beta$ be the specificity), then the number of samples that need to be assayed in the presence of any false-positive findings becomes

$$y = n\left\{\frac{1}{r} + r[1 - (1 - \beta)(1 - \pi)^r]\right\} \qquad (1)$$

The righthand columns of table 1 provide optimal pool sizes and sample size reductions under the assumption of a 5% false-positive rate for various gene frequencies.

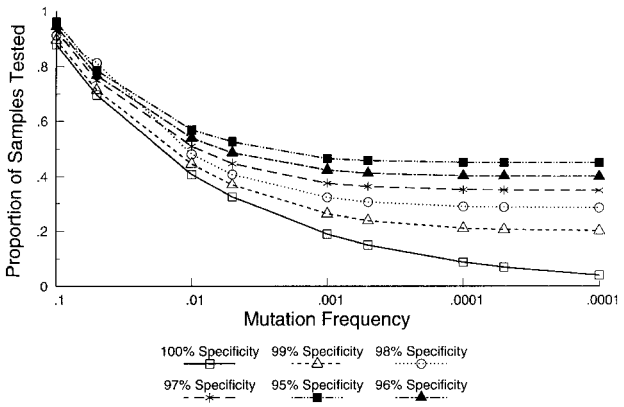

Figure 1 Proportion of tests required when pooled samples are used, with varying specificity of the test.

We have also plotted, in figure 1, the reduction in the number of tests that can be accomplished versus the mutation frequency for a range of specificities of the test. The frequency of false-positive results is the critical factor in determining the optimal pool size for any pooling strategy.

Currently available technology for sequencing restricts the pool size that can be reliably used for mutation detection. To evaluate the sensitivity of sequencing methods, we created pools of samples of DNA from nine patients previously identified to have a mutations or polymorphisms in LKB1 and who had Peutz-Jeghers syndrome (PJS). Two patients with PJS had microdeletions, and the remaining patients had missense mutations or polymorphisms; two subjects had two polymorphisms each, so that we studied 11 mutations in all. The pool size that we assayed included two or three subjects in each pool. PCR products were generated for specific exons known to have a mutation or polymorphism, and, in the pooling experiments, equal quantities of PCR product were pooled with the PCR product of the known mutation carrier or individual with a polymorphism. The samples were subjected to nucleotide-sequence analysis using an Applied Biosystems model 377 sequencer. We used the BigDye Terminator kit to minimize errors in base calling. The microdeletions were easily identifiable in pools of size two or three, and results are not further described.

To quantitate results, we used PHRED (Ewing et al. 1998). We evaluated the area under the base having the second largest area (corresponding to the mutation or polymorphism). As a referent, we quantitated the second largest areas from 6,041 bases not containing mutations derived from the original sequence analyses on the unpooled samples that identified the nine mutations. For all analyses, we deleted the first 20 bases at the beginning and end of each run, because of the increased error rates in base calling in these regions (Ewing and Green 1998). The results are shown in figure 2. Of the 6,041 referent bases, 89 had nonzero values. Compared with the pools of size two, none of the referent bases exceeded the area under the second base for the pooled sample, but, for pools of size three, 15 referent bases exceeded the area under the second base of the pooled sample. Since the largest referent bases were randomly distributed among exons, the specificity of the test (which depends upon the chance that an entire exon has a base exceeding the threshold) would have to be set at an excessively low level. Thus, pools of size three are not practical with this current technology in this study, while pools of size two were practical in this study.

The sequencing software that we have used in this preliminary study is not optimized for mutation detection in pooled samples, and so it is not surprising that the accuracy of the assay in pools of size three is not acceptable. Further development of sequencing software should allow signal identification for larger pools of samples, but even with pooling of two individuals per sample a substantial reduction in the number of tests can be accomplished. For example, application of formula 1 indicates that—for allele frequencies of .001, .0005, and .0001—only 52%, 50.4%, and 50% of the number of tests without pooling, respectively, need to be performed if the specificity of the test is 100%.

A larger set of experiments needs to be performed. Perusal of the larger second-base areas shows that the pattern of the second peak has a different appearance when it is likely to be due to background, as opposed to when it reflects a mutation in a pooled sample. In the former case, the background peaks are asymmetrical and reflect shoulders from a neighboring base, whereas, in

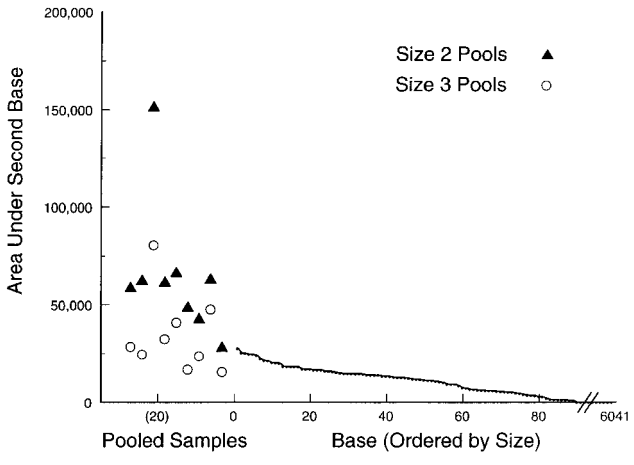

Figure 2 Distribution of area under second base for 6,041 nucleotides compared with nine sets of pooled samples. Results from different mutations are arranged in pairs, according to whether the pool size included two or three samples.

the latter case, the peak is symmetrical and may not be related to neighboring bases. We used the area under the second base to detect mutations; however, it may be possible, with a large set of sequencing runs, to develop algorithms on the basis of the area under the first base. The area under the first base is highly variable among locations and depends on the neighboring bases. However, across runs, the area under the first base shows little variation. Thus, if a substantial pool of sequences of the same exon have been completed, it should be possible to develop algorithms directed at detecting a significantly decreased signal in the first base which results when a second base is present at the same location.

This analysis neglects further approaches that might be taken to optimize mutation detection. We assume that only four colors are available for colorimetry, so that samples must be identically labeled. If individual samples could be identified within a pool by using different fluorescein dyes, further increases in efficiency would be possible. In this article, we have assumed negligible false-negative findings. For clinical studies, false-negative findings are not acceptable, and, in general, maintaining a high sensitivity may impose a bound on the number of samples that can be pooled. For population-based studies in which false-positive and false-negative findings may be acceptable, further studies should be performed to identify optimal pooling strategies, allowing for the cost to the study associated with false-positive or false-negative test results. However, our results indicate that, for assays such as sequence analysis—even when only a few samples can be pooled to yield a sensitive test—a substantial reduction in the number of tests and, therefore, a reduction in cost should be possible. Equation (1) indicates the reduction in the number of tests associated with pooling for any pool size. The efficacy of pooling is strongly related to the specificity of the test. However, even with a 5% false-positive rate, only about half as many tests would need to be performed if a pooling strategy is adopted.

Acknowledgments

We would like to thank the two anonymous reviewers and Dr. Peter Byers for their insightful comments during the review process. We thank Cynthia Thomas for assistance in manuscript preparation. Our studies have been supported by NIH grants R01ES09912, R01CA70759, P01CA34936, and R01HG52709 and by American Cancer Society grant RPG-99-030-0.

References

Barillot E, Lacroix B, Cohen D (1991) Theoretical analysis of library screening using a N-dimensional pooling strategy. Nucleic Acids Res 19:6241–6247

Bruno WJ, Knill E, Balding DJ, Bruce DC, Doggett NA, Sawhill WW, Stallings RL, et al (1995) Efficient pooling designs for library screening. Genomics 26:21–30

Chen Z-Y, Zarbl H (1997) A non-radioactive, allele specific polymerase chain reaction for reproducible detection of rare mutations in large amounts of genomic DNA: application to human K-ras. Anal Biochem 244: 191–194

Churchill GA, Giovanni JJ, Tanksley SD (1993) Pooled-sampling makes high resolution mapping practical with DNA markers. Proc Natl Acad Sci USA 90:16–20

Claus EB, Risch N, Thompson WD (1991) Genetic analysis of breast cancer in the cancer and steroid hormone study. Am J Hum Genet 48: 232–242

Coolbaugh-Murphy M, Maleki A, Strong L, Lynch P, Frazier M, Monckton D, Brown B, et al (1999) Microsatellite instability (MSI) in vitro vs. in vivo? Am J Hum Genet Suppl 65:A123

Daniels J, Holmans P, Williams N, Turic D, McGuffin P, Plomin R, Owne JM (1998) A simple method for analyzing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. Am J Hum Genet 62:1189–1197

Ewing B, Green P (1998) Base-calling and automated sequencer traces using PHRED. I. Error Probabilities. Genome Res 8:186–194

Ewing B, Hillier LD, Wendl MC, Green P (1998) Base-calling and automated sequencer traces using PHRED. I. Accuracy assessment. Genome Res 8:175–185

Hartge P, Struewing JP, Wacholder S, Brody LC, Tucker MA (1999) The prevalence of common BRCA1 and BRCA2 mutations among Ashkenazi Jews. Am J Hum Genet 64: 963–970

Krook A, Stratton IM, O'Rahilly S (1992) Rapid and simultaneous detection of multiple mutations by pooled and multiplex single nucleotide primer extension: application to the study of insulin-responsive glucose transporter and insulin receptor mutations in non-insulin-dependent diabetes. Hum Mol Genet 1:391–395

Mohrenweiser HW, Jones IM (1998) Variation in DNA repair is a factor in cancer susceptibility: a paradigm for the promises and perils of individual and population risk estimation? Mutat Res 400:15–24

Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A (1998) Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. Genome Res 8:111–123

Sheffield VC, Nishimura DA, Stone EM (1995) Novel approaches to linkage mapping. Curr Opin Genet Dev 5: 335–341

Taillon-Miller P, Piernot EE, Kwok P-Y (1999) Efficient approach to unique single-nucleotide polymorphism discovery. Genome Res 9:499–505